OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE

Finding Patterns of Emergence in Science and Technology

LEADING INTELLIGENCE INTEGRATION

Dewey Murdick, Program Manager
Office of Incisive Analysis, IARPA
24 September 2012

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
24 SEP 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Foresight and Understanding from Scientific Exposition (FUSE)	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Office of Incisive Analysis, IARPA,Washington,DC, 20511**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the Foresight and Understanding from Scientific Exposition (FUSE), Washington, DC, 24 September 2012

14. ABSTRACT

15. SUBJECT TERMS

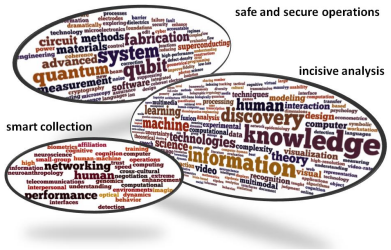
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**40**	

IARPA

BE THE FUTURE

"Invests in high-risk/high-payoff research programs that have the potential to provide our nation with an overwhelming intelligence advantage"

http://www.iarpa.gov/



Goal: Validated, early detection of technical emergence

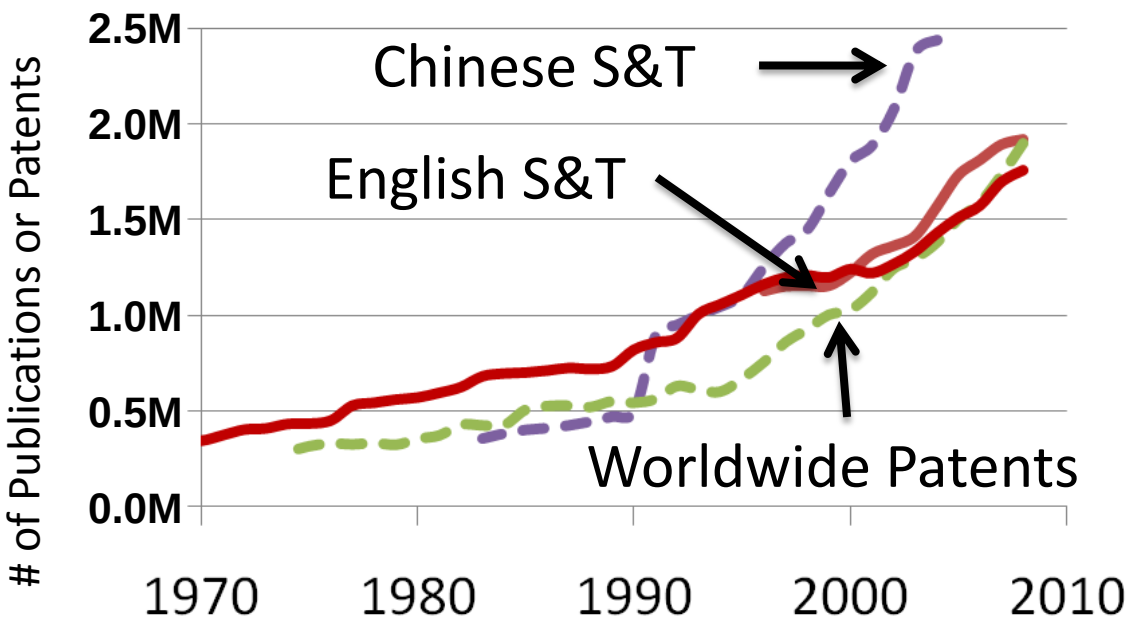
Enable reliable, early detection of emerging scientific and technical capabilities across disciplines and languages found within the full-text content of scientific, technical, and patent literature

Focus from the outset on **English, Chinese, German**, Japanese, Russian, *Korean*, and *Spanish*

Novelty	→	Discover patterns of emergence and connections between technical concepts at a speed, scale, and comprehensiveness that exceeds human capacity
Usage	→	Alert analyst of emerging technical areas with sufficient explanatory evidence to support further exploration

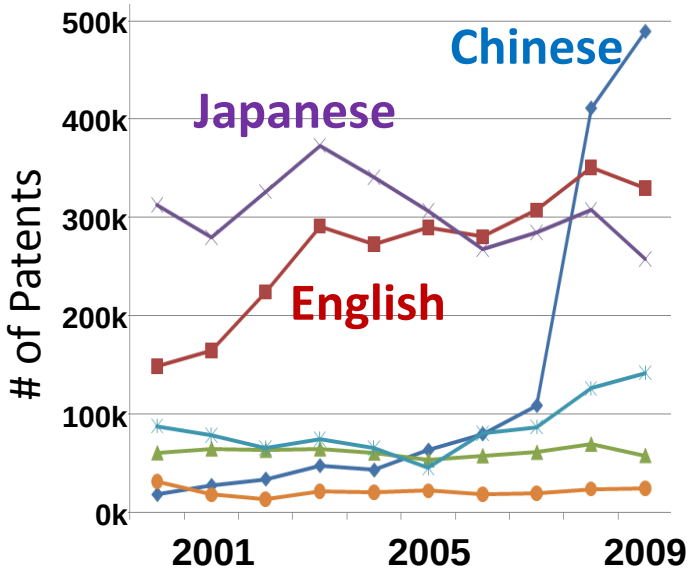
Worldwide Scientific and Patent Literature

Publications (by Language) and Patents



Growth in scientific and patent literature is estimated at 800k docs/month

Patents by Language



Other Languages

Korean (cyan)
German (green)
Russian (orange)

FUSE Approach

Today, *ad hoc* "technical horizon scanning" already consumes substantial expert time, is narrowly focused on a small number of topics, and is subject to limited systematic validation.

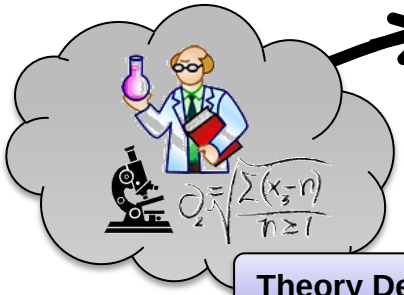
Analysts need to <u>scan continually</u> for signs of technical capability emergence.

Today	FUSE
Manual	Automatic
Selected coverage	"Complete" literature coverage
Updated infrequently	Updated monthly
Months to produce (for one technical area)	24hrs to produce (for all technical areas)
Ad hoc evaluation	Formal models of emergence

Complete, Continual, Unbiased

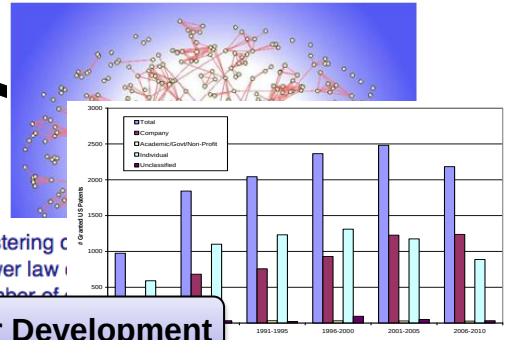
FUSE Research Thrusts

All teams are pursuing all research areas in parallel



Theory Development

Document Features



- clustering
- power law

Indicator Development

Nomination Quality

Challenge Question Nomination

RNAi: *Community of Practice*					
1996 – 2000		2001 – 2005		2006 - 2010	
Y	.54	Y	.82	Y	.96
RNAi: *Practical Application*					
1996 – 2000		2001 – 2005		2006 - 2010	
N	.81	Y	.77	Y	.98

Evidence Representation

RNAi : 2006-2010 : CQ1

Was there a community of practice around RNAi during 2006-2010?

The answer is YES, with a confidence of 72%

Many indicators suggest a positive answer to the CQ, especially within the Coaul

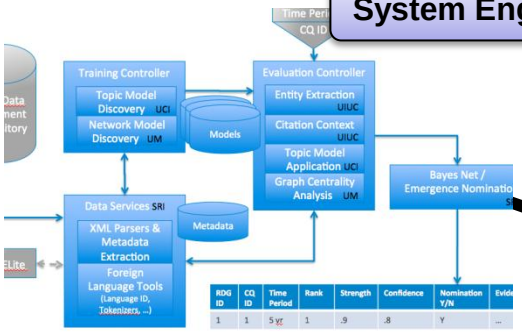
Coauthorship Graph

Click for detailed view

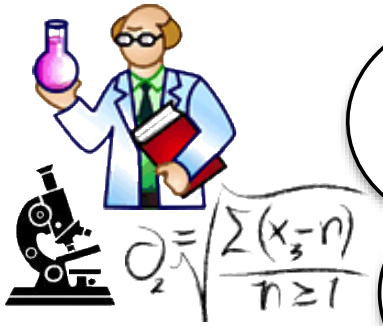
The coauthorship graph for RNAi spans 520 authors, and it has the properties of a small-v communities. It is a fully connected network with a high clustering coefficient as well as a

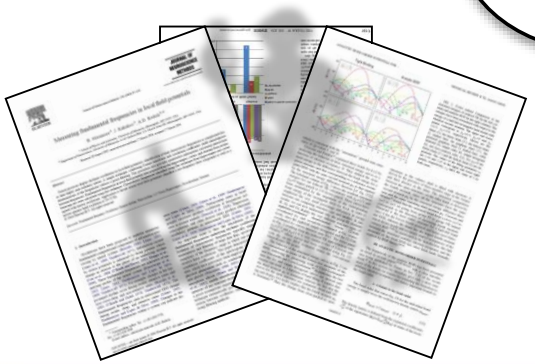
Coauthorship graph indicators are the most powerful when determining the answer for CQ the direction of a positive answer.

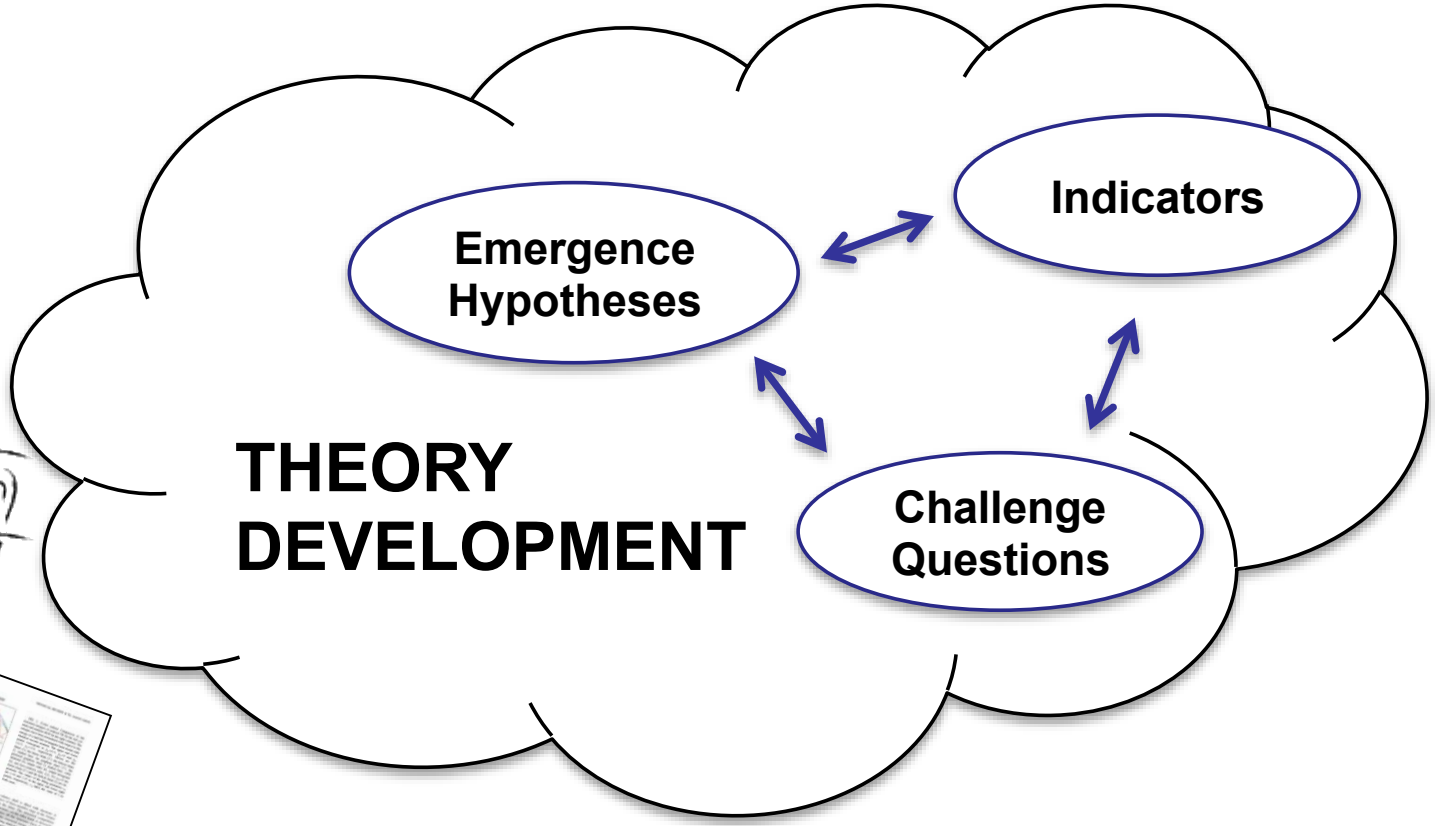
System Engineering



FUSE Theory Exploration







What is technical emergence?

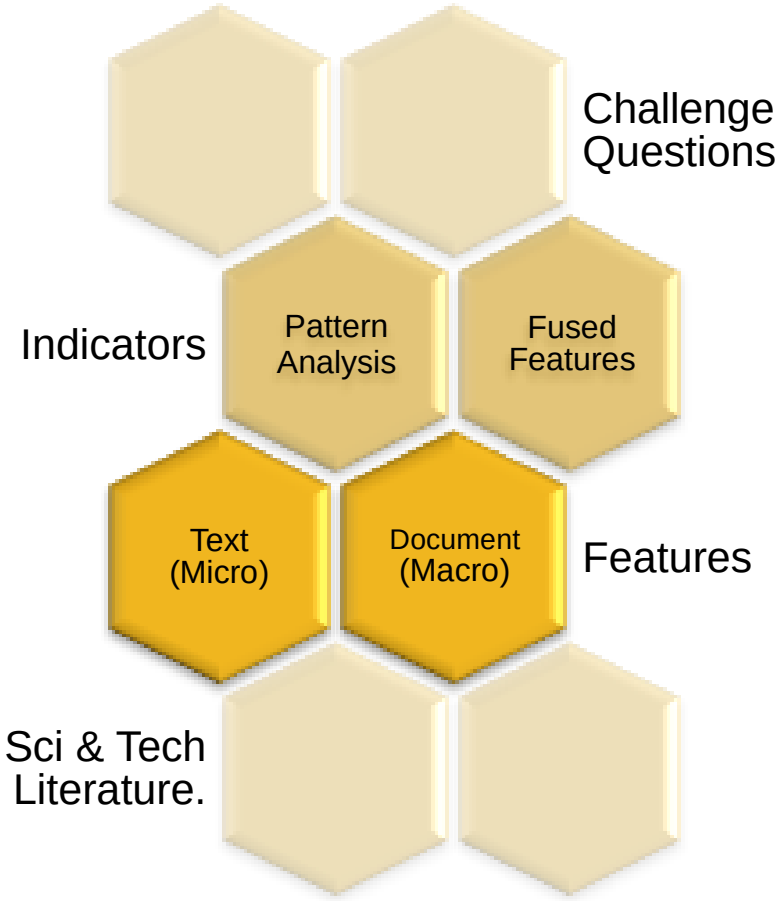
Current Hypotheses

- A concept has emerged if it has been accepted by others within and beyond one's community. **~Columbia**
- A concept is emerging when its actant network is increasing in robustness. **~BAE**
- A concept has emerged when evidence has appeared that the concept is new and unexpected, noticeable and growing. **~Raytheon BBN**
- A concept is emerging when it is identifiable by its own practitioners, enables a capability that was not achievable previously, and persists. **~SRI**

Many ways to probe technical emergence

- Community of Practice
- Practical Application
- Debates
- Alternative
- Acceptance
- Interdisciplinarity
- Attention (Citation) Prediction
- Dominant sub-topic within set
- Commercial Application
- Infrastructure

Text to Indicator Implementation



New classes of FUSE-related features are being developed and validated

- Community response, role in community
- Topic stability
- Rhetorical stance for refs/citations
- Zoning the full-text content
- Methods moving from “the topic” to “used” to “mentioned”
- Quality and quantity of resources available to an investigator or inventor
- New terms introduction and adoption

Nomination and Evidence Explanation

RNAi : 2006-2010 : CQ1

Was there a community of practice around RNAi during 2006-2010?

The answer is YES, with a confidence of 72%

Topic Summary

RNA interference (RNAi) is an RNA-dependent gene silencing process within living cells. The selective and robust effect of RNAi on gene expression makes it a valuable research tool, both in cell culture and in living organisms.

Ongoing Work

Justification and evidence for answer

Many indicators suggest a positive answer to the C
Coauthorship Graph, Time Series and Funding gr

Coauthorship Graph

Click for detailed view

The coauthorship graph for RNAi spans 520 authors, and i
small-world network which is typical of real-world communi
network with a high clustering coefficient as well as a reas
coefficient.

Coauthorship graph indicators are the most powerful when
CQ1, and in this case their values strongly point in the dire

Time Series

Click for detailed view

Time series

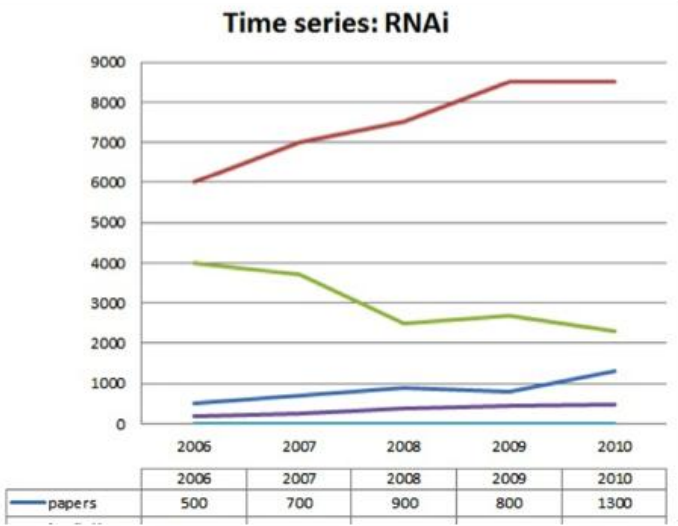
A time series is a sequence of data points measured at successive time instants spaced at uniform time intervals (in our case, years). Time series analysis looks at the way that various functions of the RDG behave over time.

The time series for number of papers and number of unique authors have very high slopes, indicating a lively community of practice which evolves over time. The number of in-citations and out-citations also progress visibly over time.

Time series indicators are very important when determining the answer for CQ1. In this case, they clearly suggest a positive answer.

Indicators:

- slope, number of papers: 170
- slope, number of in-citations: 650
- slope, number of out-citations: -440
- slope, number of authors: 78.1
- slope, number of conferences: 0.2



	2006	2007	2008	2009	2010
papers	500	700	900	800	1300

FUSE Validation / Metrics

- Guiding the development of theories and indicators of emergence
  - *"Emergence Theory Workshop and Peer Review"*
- Effective identification, prioritization and nomination of technical areas as compared to real world (e.g., experts, case studies, present day tests for both positive / negative examples)
  - *"Nomination Quality"*
- Evidence provided in a clear and humanly usable form
  - *"Evidence Quality"*
- System to perform at scale across multiple languages
  - *"Computational Efficiency" and "Multilingual Performance"*

- Measure technical emergence from "real world" viewpoint that is drawn from diverse areas of scientific inquiry & application
  - *"Case studies and reference baseline"*

FUSEnet – Computational Environment

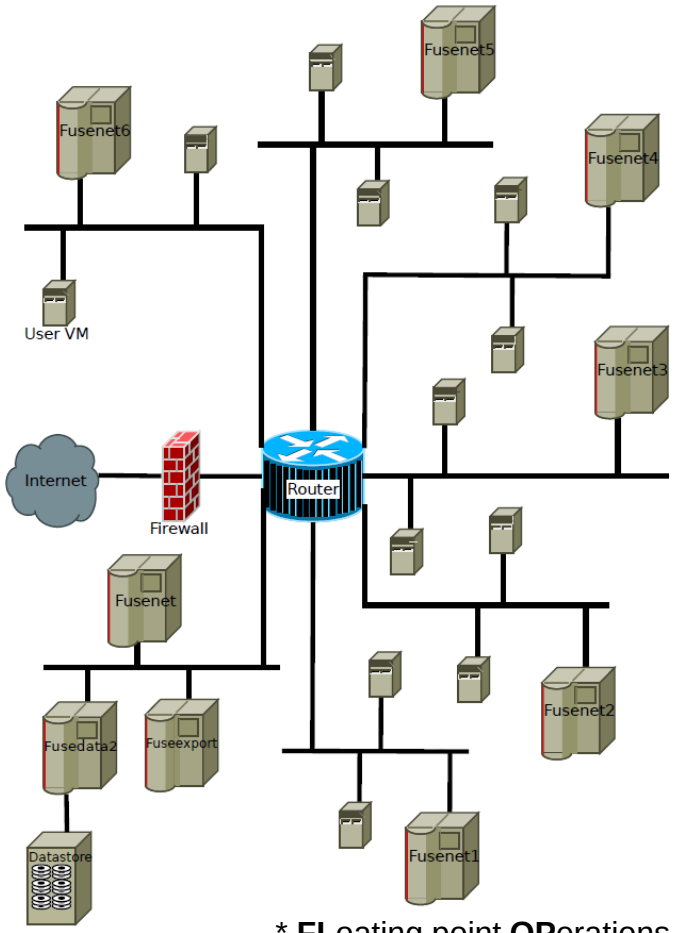
FUSEnet

- Government system hosted by Oak Ridge National Laboratory (ORNL)
- Protected unclassified system with remote access for all approved users

Current Specifications for FUSEnet 1.0

- 770 gigaFLOPS* of maximum performance
- 16 blade servers, each with 6 cores, totaling 192 processors
- 2 blade servers for backup (not in FLOPS estimate)
- 96 GBytes of RAM per server for a total of 1.5 TBytes
- 260 TBytes of effective disk storage
- iSCSI 10 Gigabit connectivity
- Virtualized computing space through VMware
- Access to Document Repository (DR) through iSCSI
- Access and control policies are enforced by ORNL
- Call Center and metrics for service quality

FUSEnet 2.0 Specifications Pending



* **FL**oating point **OP**erations per **S**econd

If you had a system that could

(a) reliably identify what technical capabilities are emerging
and
(b) provide humanly understandable evidence explanation
then

how would you use this tool?

How might an IARPA PM use FUSE?

- **Idea development for a relevant problem or problem space**
  - Answer Heilmeier Questions (http://www.iarpa.gov/join3.html)
  - Hype versus reality
  - New enabling component capabilities, signs of potential convergence
  - Why is this innovative? Does it overlap with past and existing efforts?
  - Does new capability or convergence of capabilities require further investment to motivate progress? Optimal emergence state for an R&D program investment
- **Program impact assessment**
  - Observable impact within a research community, across research communities, beyond the research community?
    - Are these focused programs yielding novel results or not?
    - Is application development potential increasing as a result of this effort?
    - Is there a gap between funding and output?
  - Which areas are being driven by a focused research program? Which ones are not (but still have a vibrant community of practice)? Which ones are being driven by many research programs?

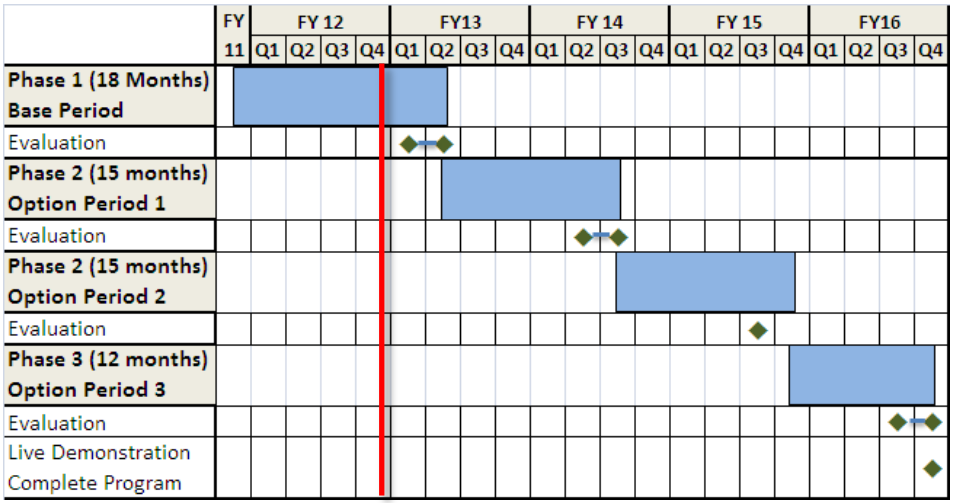
Would the components of FUSE enable a new capability or enhance an existing workflow?

Potential Component Services

Lineage / Provenance of Acquired Data	Data Transformation (e.g., PDF to XML)	XML Enrichment (e.g., new extractions, zoning)
Cross-Document "Feature" Linkage	Generation of Document Groupings	Indicators (e.g., Confirming, Dis-confirming)
Nomination, Prioritization by Time Period	Evidence Explanation and Justification for Nomination	

FUSE Status Update

- Five year, fundamental research program
- Teams under contract since August 2011
  - BAE Systems
  - Columbia University
  - Raytheon BBN
  - SRI International



- Formal test and evaluation to begin October 2012, three additional rounds of formal evaluation scheduled
  - Case Studies, Eight Examples: Tissue Engineering, Cold Fusion, RF Metamaterials, DNA Microarrays, Genetic Algorithms, RNAi, Steganography, Horizontal Gene Transfer

The FUSE Team

- 3 (+2) large businesses
- 3 (+3) small businesses
- 14 academic orgs
- 1 not-for-profit org
- Many data vendors
- Plus FFRDCs & gov orgs

A Few Unique Qualities of FUSE

- **Connecting disparate communities with lots of data sows the seeds for much discovery**
  - Social science, emergence theory, Natural Language Processing (NLP), etc.
- **Scientific, Technical, and Patent literatures are new genres**
  - Most Natural Language Processing (NLP) trained on news wire
  - Foreign language tokenization, etc., too
- **Novel extrinsic test for technical emergence**
  - Measuring detection of emergence (S&T)
  - Not measuring NLP metric X *(internal measures)*
- **Challenging scale, but tractable**
  - Concept extraction, within/cross-doc linkages
  - Currently growing on order 800k/mo. (sci lit + patents), FUSE has ~75-80% filed patents & ~10% sci lit (1980-2010)
- **Automated evidence explanation is really challenging, but exciting**
- **Synthesize & Scan "Horizon," not Search**

Anticipated Impact

- **Scientific & Technical Analysis Impact**
  - Relevant, timely, and bias-controlled analytic force multiplier to maintain technical vigilance, across all disciplines and multiple languages
  - Discover previously unknown emergence signals of interest at speed, scale, and comprehensiveness that exceeds human capacity
- **Technical Impact**
  - Generalized and validated theories of technical emergence
  - New cross-document conceptual feature extraction technologies
  - Progress in computer-generated evidence representations for human use
- **Secondary Impact**
  - Improved priority filter for USG investment strategies and policy

Questions

Dewey Murdick, Ph.D.
FUSE Program Manager, IARPA
dewey.murdick@iarpa.gov

Finding patterns of emergence in science and technology

Abstract

Today, the identification and assessment of emerging technical capabilities is a time-consuming, domain-specific, and expert-intensive process. This demanding process is often carried out under severe time constraints on either too much or too little data, with limited reproducible auditing and bias controls, and with limited systematic validation against real world activities. Furthermore, the increasing globalization of science and technology raises the potential for high-impact technical capabilities to emerge in increasingly diverse technical, socio-economic, and geographic areas.

Analysts, subject-matter experts, and even research program managers benefit from a reliable, evidence-based capability that allows them to dramatically accelerate the horizon-scanning process and reduce the labor involved to identify specific emerging technical areas in context for in-depth review. The Foresight and Understanding from Scientific Exposition (FUSE) Program is the Intelligence Advanced Research Projects Activity (IARPA) response to this need.

The FUSE Program seeks fundamental advances in our understanding of how the real-world processes of technical emergence leave discernible traces in the public scientific, technical, and patent literature, and how those traces can be detected, fused, and prioritized. FUSE aims to develop and validate a comprehensive suite of quantitative measures of technical emergence that generalize across language, culture and technical area. Technology developed from the FUSE Program will automatically nominate both known and novel technical areas based on quantified indications of technical emergence with sufficient supporting evidence and arguments for that nomination.

The presentation will introduce the technical approach and explore the potential impact of technologies and insights that may emerge as a result of the FUSE Program. For more information, see http://www.iarpa.gov/solicitations_fuse.html.

BACKUP

Foresight and Understanding from Scientific Exposition (FUSE) Program

GOAL: Enable analysis by reliably detecting emerging scientific and technical capabilities across disciplines and languages found within the full-text content of scientific, technical, and patent literature (in EN, CN, DE, JP, RU, …)
WHY: Analysts need effective ways to maintain technical vigilance and discover previously unknown capabilities in a relevant, timely, and bias-controlled analytic manner
HOW: Develop theories and indicators of technical emergence; process full-text literature for relevant features; identify, prioritize & nominate high-priority technical areas and provide evidence with understandable explanations for analytic use
SUCCESS: Systems that "speed read" 100s of millions of pages of technical text and provide understandable and useful evidence that justifies high-priority alerts of emerging capabilities

Case Studies, Eight Examples:
- Tissue Engineering, Cold Fusion, RF Metamaterials, DNA Microarrays, Genetic Algorithms, RNAi, Steganography, Horizontal Gene Transfer

Accomplishments:
- Currently in Phase 1 through 7 Feb 2013
- Four research teams developing, testing, and validating theories of emergence, indicator efficacy, nomination precision and recall, evidence clarity, and scalable system performance

Upcoming milestones:
- July 2012 intermediate system test
- Oct 2012 - Jan 2013 system test and evaluation cycle

Specific Phase 1 Goals

- Craft viable theories / hypotheses of **technical emergence**
- Effectively use **full-text features** and measure impact on technical emergence
- Correctly **nominate document groups** that exhibit technical emergence as represented by challenge questions, time periods
  - Valid nomination extends across disciplines
  - Satisfactory prioritization of topics over time periods
- Establish a reliable measure for **Evidence Quality** (i.e., the rubric) and deliver comprehensible evidential support for nomination
- Demonstrate **proof-of-concept nomination for Chinese and German** topics with realistic progress on multilingual components
- Demonstrate system functionality that establishes confidence that team can **transition to Phase 2** (e.g., scalability, minimize brittleness across disciplines and document types)

Currently 1 year into a 5 year program … parallel efforts at basic research components and prototyping / system engineering.

Why now?

- Important problems to overcome:
  - Need to learn to automatically ***scan*** for emergence (beyond search)
  - Too much information to analyze, in too many languages
    - Support strategic investment
    - Facilitate discovery and innovation
  - Cannot reliably query for patterns that indicate emergence without starting with a known, named subject
- Automated analysis is likely to work because:
  - The scientific literature is now available in digital formats
    - Metadata records are well curated and ready for use
    - Exploitation of the full text of documents is now possible (although not easy)
  - Emerging text and “signal” analysis (temporal pattern) techniques are promising
    - Context-sensitive feature extraction from text
    - Unsupervised clustering
    - Machine learning
    - Statistical modeling
    - Pattern matching and analysis
    - Indicator development and validation

Program Structure

Phase (Period)	Duration	Primary English and *Multilingual* Goals
Phase 1 (Base Period)	Aug 2011 – Feb 2013 (18 mo)	**Demonstrate that full-text literature can be the source for robust indicators of technical emergence within a consistent theoretical construct.** Automatically prioritize a small number of provided Related Document Groups (RDGs), each representing a single technical area. Nominate those RDGs that exhibit technical emergence. *Demonstrate proof-of-concept functionality in at least two languages in addition to English.*
Phase 2 (Option Periods 1 & 2)	Feb 2013 – May 2014 (15 mo) May 2014 – Aug 2015 (15 mo)	**Demonstrate automatic generation and nomination of those RDGs that exhibit single technical area emergence**, from a collection of millions of full-text documents. *For at least two languages in addition to English, automatically prioritize provided RDGs, each representing a single technical area. Nominate those RDGs that exhibit technical emergence.*
Phase 3 (Option Period 3)	Aug 2015 – Aug 2016 (12 mo)	Demonstrate automatic generation and nomination of those RDGs that exhibit **technical emergence across disparate technical areas**, from a collection of millions of full-text documents. *For at least two languages in addition to English, demonstrate automatic generation and nomination of those RDGs that exhibit single technical area emergence, from a collection of full-text documents.*

Data Files Loaded on the FUSEnet "DR" Per Source

Data Source (update)	Size (GB)	# Metadata XML Files	# Full-text XML Files	# Full-Text ETL XML Files	# Full-text PDF Files	# Image Files
Elsevier (yr)	264		3,825,122			
IEEE-ASPP (opt)	393	383,212		491,232	483,102	
IEEE-POP (opt)	1,136	1,908,581			1,908,417	
IEEE-TJ (opt)	434		72,968		92,788	
Lexis-Nexis – CN	442		5,122,175			4,007,340
LN – DE	235		3,896,793			1,618,995
LN – EP	427		4,213,461			1,456,570
LN – GB	79		1,656,114			?
LN – JP	918		20,897,653			9,306,387
LN – KR	749		4,310,680			
LN – RU	36		696,841			64,045
LN – SU	23		874,772			21
LN – US	1,100		7,101,711			5,592,447
LN – WO	363		2,546,359			?
LN All Update (mo)	3,507		58,425,015			5,173,584
Nature	6		199,083			
PUBMED Central+ (opt)	728		229,393		253,622	
Scopus (yr)	685	37,642,585				
SPIE	10	16,877			18,173	
Web of Sci (6 mo)	574	48,693,056				
Total	**12,528**	**88,627,434**	**100,886,913**	**738,860**	**2,756,102**	**26,095,450**

How do we probe "Technical Emergence"?

Government-Defined Challenge Questions

- Was there a **community of practice** around <concept> during <time period>?
- Were there **debates** within the scientific community on <concept> during <time period>?
- Was there a demonstration of **practical application** of <concept> during <time period>?
- Was <concept> considered an **alternative** to an established concept during <time period>?

- *Was there a demonstration of commercial application of <concept> during <time period>?*
- *Was the infrastructure required to perform research in <concept> readily available during <time period>?*

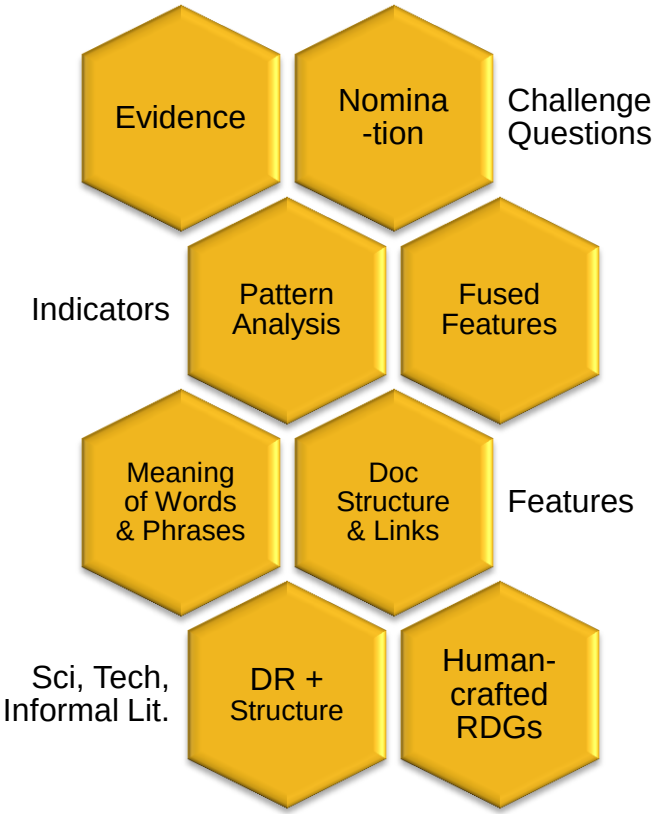
Performer-Defined Challenge Questions

- Was <concept> accepted during <time period>? ***Columbia***
- Did the acceptance of <concept> increase or decrease during <time period>? ***Columbia***
- How interdisciplinary was the scientific and technical knowledgebase around <concept> during <time period>? ***SRI***
- Did usage of new terminology describing <concept> increase in robustness during <time period>? ***BAE***
- How many citations of papers from <concept> published in <time period 1> would you expect to see in <time period 2>? ***Raytheon BBN***
- Does the <concept> dominate a thread during <time period>? ***Raytheon BBN***

Case Studies

- Drawn from diverse areas of scientific inquiry & application:
  - Biological Sciences / Biotechnology
  - Computer Science / Information Science; Engineering
  - Mathematics / Statistics
  - Physical Sciences; Earth Science
  - Medical / Clinical / Infectious Disease / Health Services;
  - Social Sciences; …
- Technical emergence measured from “real world” view point, but connected to literature
- Multiple case studies to be produced; some are held back for evaluation
  - Case studies are representative but not comprehensive
  - Insufficient to train technical emergence classifiers
  - Limited examples of emergence & non-emergence (10s planned)
  - Reference baseline has limited temporal resolution (~5 year blocks)

FUSE Horizon Scanning



Emergence Queue(s) for Analysis		
Community of Practice	***Practical Application***	***Performer CQ***
1. RDG # [why?]	1. RDG # [why?]	1. RDG # [why?]
2. RDG # [why?]	2. RDG # [why?]	2. RDG # [why?]
…	…	…

Generic Use Case (e.g., Alerts and Interactive Queue Exploration and Analysis)

[why?] → Eval Interface (e.g., Web-Browser)
- Descriptive Text, Indicator template, …
- Time Series, Network Graphs, …
- Text Snippet, Generated Text, Doc Refs, …

IARPA

IARPA Overview

IARPA's mission is to invest in high-risk/high-payoff research programs that have the potential to provide the U.S. with an overwhelming intelligence advantage over our future adversaries

- CAVEAT: HIGH-RISK/HIGH-PAYOFF IS NOT A FREE PASS FOR STUPIDITY.
- Bring the best minds to bear on our problems.
  - World-class Program Managers (PMs).
    - IARPA will not start a program without a good idea and an exceptional person to lead its execution.
  - Full and open competition to the greatest possible extent.
- Cross-community focus.
  - Address cross-community challenges & leverage community expertise
  - Work transition strategies and plans

http://www.iarpa.gov

IARPA Offices and Areas of Emphasis

safe and secure operations

incisive analysis

smart collection

The "Heilmeier Questions"

1. What are you trying to do?
2. How does this get done at present? Who does it? What are the limitations of the present approaches?
   - Are you aware of the state-of-the-art and have you thoroughly thought through all the options?
3. What is new about your approach? Why do you think you can be successful at this time?
   - Given that you've provided clear answers to 1 & 2, have you created a compelling option?
   - What does first-order analysis of your approach reveal?
4. If you succeed, what difference will it make?
   - Why should we care?
5. How long will it take? How much will it cost? What are your mid-term and final exams?
   - What is your program plan? How will you measure progress? What are your milestones/metrics? What is your transition strategy?

The "P" in IARPA is very important

- Technical and programmatic excellence are required
- Each Program will have a clearly defined and measurable end-goal, typically 3-5 years out.
  - Intermediate milestones to measure progress are also required
  - Every Program has a beginning and an end
  - A new program may be started that builds upon what has been accomplished in a previous program, but that new program must compete against all other new programs
- This approach, coupled with rotational PM positions, ensures that…
  - IARPA does not "institutionalize" programs
  - Fresh ideas and perspectives are always coming in
  - Status quo is always questioned
  - Only the best ideas are pursued, and only the best performers are funded.

Office of Incisive Analysis

"maximizing insight from the information we collect, in a timely fashion"

Large Data Volumes and Varieties	Social-Cultural and Linguistic Factors	Improving Analytic Processes
Providing powerful new sources of information from massive, noisy data that currently overwhelm analysts.	Analyzing language and speech to produce insights into groups and organizations.	Dramatic enhancements to the analytic process at the individual and group level.

Office of Smart Collection

"dramatically improve the value of collected data"

Novel Sources of Information

Create innovative technologies and tools for reaching hard targets in denied areas

Identity Intelligence

- Detect the trustworthiness of others
- Advance biometrics in real-world conditions

Tracking and Locating

Accurately locate HF emitters and low-power, moving emitters with a factor of ten improvement in geolocation accuracy

Office of Safe and Secure Operations

"counter emerging adversary potential to deny our ability to operate effectively in a globally-interdependent and networked environment"

Computational Power

Revolutionary advances in science and engineering to solve problems intractable with today's computers

Trustworthy Components

Getting the benefits of leading-edge hardware and software without compromising security

Safe and Secure Systems

Safeguarding mission integrity in a hostile world